Beaconcure Table Mining

By: Orly Frank, VP R&D, Beaconcure

Background

In the life-science industry, statistical analysis reports contain hundreds and thousands of tables. A table is more than its text. The table includes a visual representation of the data that helps people understand the various numbers presented. If a company wants to save time and increase efficiency by automating the process of validating the statistical analysis, it must have its tables in a machine-readable format.

What Makes the Process of Creating the Tables in a Machine-Readable Format Challenging?

Machine-readable data is a data format that can be easily read and processed (identified and extracted) by a computer, without human intervention, including individual statements, and their internal structure, while ensuring no semantic meaning is lost. Machine-readable data must be structured data.
Presenting information in tables is very challenging for automatic processing, since the traditional text mining techniques fail to capture the meaning of the data.
Clinical data tables are usually unstructured files, designed for the human eye. For the machine to understand the numbers and values we will have to use advanced techniques of table mining.

The table below (Table 14.7) is an example of a typical clinical data table. It is clear for humans that the value '15' highlighted twice in the table has a different meaning in each cell, but how can the machine understand the difference?

Table 14.7 - Unstructured file

Table 14.7
Shift Table of Lab Data - Glucose

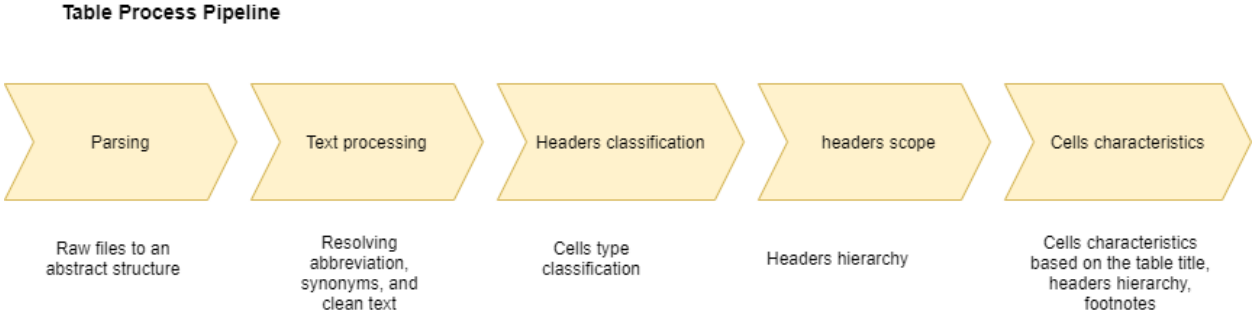
	Drug A (N=119)				
	Missing n (%)	Below Normal n (%)	Within Normal n (%)	Above Normal n (%)	Total at Visit n (%)
6 Months:					
Missing	1 (0.8%)	0	2 (1.7%)	0	3 (2.5%)
Below Normal Range	0	12 (10.1%)	0	0	12 (10.1%)
Within Normal Range	0	3 (8.4%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
Above Normal Range	0	0	12 (10.1%)	20 (16.8%)	32 (26.9%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
12 Months:					
Missing	1 (0.8%)	0	0	0	1 (0.8%)
Below Normal Range	0	10 (8.4%)	5 (4.2%)	0	15 (12.2%)
Within Normal Range	0	5 (4.2%)	65 (54.6%)	16 (13.4%)	86 (72.3%)
Above Normal Range	0	0	7 (5.9%)	10 (8.4%)	17 (14.3%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
18 Months:					
Missing	0	0	0	0	0
Below Normal Range	0	3 (5.0%)	4 (6.7%)	0	7
Within Normal Range	0	2 (3.3%)	40 (66.7%)	2 (3.3%)	44
Above Normal Range	0	0	4 (6.7%)	5 (8.3%)	9
Total at Baseline	0	5 (83.0%)	48 (80.0%)	7 (11.7%)	60 (100%)
24 Months:					
Missing	0	0	0	0	0
Below Normal Range	0	2 (3.3%)	6 (10.0%)	0	8
Within Normal Range	0	2 (3.3%)	36 (60.0%)	4 (6.7%)	42
Above Normal Range	0	1 (1.7%)	6 (10.0%)	3 (5.0%)	10
Total at Baseline	0	5 (83.0%)	48 (80.0%)	7 (11.7%)	60 (100%)

a. N = number of subjects in the specified group, or the total sample. This value is the denominator for the percentage calculations.
b. n = Number of subjects with the specified characteristic.
c. Days calculated since Dose 1.
d. Protocol-specified time frame.

Beaconcure Table Mining - Table Process Pipeline

Clinical data tables exist in a variety of formats with different styling; Beaconcure table mining standardizes the data to an abstract hierarchy repository.

The 'Table Process Pipeline' (Diagram below) converts the data from the raw unstructured file, which is unstructured, to a structured machine-readable data. Our unique solution includes ML models trained on clinical data tables for cell classification (header cells vs. data cells) and data hierarchy.



Data Representation

Table mining via the Beaconcure 'Table Process Pipeline' enables the machine to understand the data within the tables.

The value '15' appears twice in the same table and has a different meaning in the different cells. When a human eye is looking at the table, it immediately understands the context of each cell, based on its referenced headers.

In Beaconcure's database, the representation of the data cells includes additional attributes:

The corresponding column headers and row headers relate to the values that give the meaning to its data tokens.

Data representation in Beaconcure database for the value '15' data cell

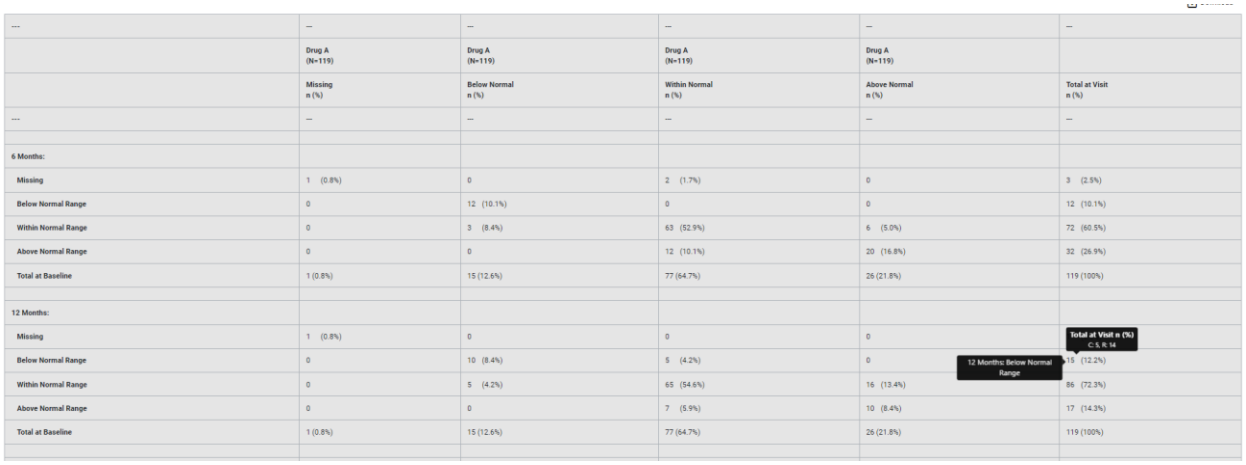

```
{
    "col_val_l": [
        "<b>Drug A</b>\n<b>(N=119)</b>",
        "<b>Below Normal</b>\n<b>n (%)</b>"
    ],
    "row_val_l": [
        "6 Months:",
        "   Total at Baseline"
    ],
    "token_l": [
        "15",
        "12.6",
        "%"
    ],

},
```


Data representation for Value '15' post Beaconcure data mining tool:

Table 14.7 - Beaconcure analysis view - understanding the value '15' in the first cell

Table 14.7 - Beaconcure analysis view - understanding the value '15' in the second cell



	Drug A (N=119)	Drug A (N=119)	Drug A (N=119)	Drug A (N=119)	
	Missing n (%)	Below Normal n (%)	Within Normal n (%)	Above Normal n (%)	Total at Visit n (%)
6 Months:					
Missing	1 (0.8%)	0	2 (1.7%)	0	3 (2.5%)
Below Normal Range	0	12 (10.1%)	0	0	12 (10.1%)
Within Normal Range	0	3 (8.4%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
Above Normal Range	0	0	12 (10.1%)	20 (16.8%)	32 (26.9%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
12 Months:					
Missing	1 (0.8%)	0	0	0	1 (0.8%)
Below Normal Range	0	10 (8.4%)	5 (4.2%)	0	15 (12.2%)
Within Normal Range	0	5 (4.2%)	65 (54.6%)	16 (13.4%)	86 (72.3%)
Above Normal Range	0	0	7 (5.9%)	10 (8.4%)	17 (14.3%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)

Data Visualization

Below is the table's new visualization that facilitates faster, easier and more accurate clinical data validation. In the near future end users will be able to validate, query, and edit the content easily.

Table 14.7
Shift Table of Lab Data - Glucose

	Drug A (N=119)				
	Missing n (%)	Below Normal n (%)	Within Normal n (%)	Above Normal n (%)	Total at Visit n (%)
6 Months:					
Missing	1 (0.8%)	0	2 (1.7%)	0	3 (2.5%)
Below Normal Range	0	12 (10.1%)	0	0	12 (10.1%)
Within Normal Range	0	3 (8.4%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
Above Normal Range	0	0	12 (10.1%)	20 (16.8%)	32 (26.9%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
12 Months:					
Missing	1 (0.8%)	0	0	0	1 (0.8%)
Below Normal Range	0	10 (8.4%)	5 (4.2%)	0	15 (12.2%)
Within Normal Range	0	5 (4.2%)	65 (54.6%)	16 (13.4%)	86 (72.3%)
Above Normal Range	0	0	7 (5.9%)	10 (8.4%)	17 (14.3%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)

Beaconcure visualization of the data enables discrepancies (in the tables) to be more accurately and effectively identified, understood and tracked.

To learn more about data visualization for clinical data validation you can read a post we published on our blog:

https://beaconcure.com/data-visualization-for-clinical-data-validation/

Conclusion

Due to the diverse nature of clinical trial data, it remains a challenge to harness this data for analytics. Beaconcure's table mining algorithms provide the life-science companies better accessibility to their data and the ability to apply different kinds of applications, once the data is structured in a machine-readable format. Furthermore, Beaconcure analyzes the inner grouping of the table, and the relations between them. Table mining is the core of our capabilities that allows us to perform all checks comprehensively and consistently to ensure full validation for drugs and vaccines.

For more information on implementing a state-of-the-art solution for statistical analysis validation of clinical trials outputs, visit Beaconcure today at https://beaconcure.com/ hello@beaconcure.com

Copyright 2021 Beaconcure, Orly Frank, all rights reserved